ARTIFICIAL INTELLIGENCE IN
MEDICINE

VOL. 37, ISSUE 1 MAY 2006

CONTENTS

Abstracted/Indexed in: Biomedical Engineering Citation Index; Cambridge Scientific Abstracts; Computer Abstracts; Current Contents, Clinical Medicine; EMBASE; Engineering Index/Compendex; Index Medicus/MEDLINE; INSPEC Information Services; Science Citation Index; Sci Search. Listed as an approved journal by the Publications Committee of the American Association for Artificial Intelligence.

www.intl.elsevierhealth.com/journals/aiim

05095

37
1

Artificial Intelligence in Medicine Vol. 37/1 (2006) 1–72 Special Issue: Intelligent Data Analysis in Medicine

ELSEVIER

Volume 37 No 1 May 2006 ISSN 0933-3657

SPECIAL ISSUE:
Intelligent Data Analysis in Medicine

Guest Editor:
P. Perner

Artificial Intelligence in Medicine (2006) 37, 1–5

ELSEVIER

ARTIFICIAL
INTELLIGENCE
IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

GUEST EDITORIAL

Intelligent data analysis in medicine—Recent advances

1. Introduction

Medical and biomedical intelligent data analysis is a complex and very important field. Although a lot of work has been done based on statistical methods, there has been little progress and the medical doctors admit that they are still doing evidential medicine instead of making diagnoses based on hard facts. There is still a lot to do in order to introduce methods that can help physicians make their diagnoses based on objective data and methods. Medical and biomedical intelligent data analysis is still a challenging and inspiring field because of the diversity of the data and the special circumstances in which the data are acquired. We are reviewing in this paper the progress in the field based on the papers that were selected for this special issue and other papers presented at the Industrial Conference in Data Mining ICDM-Leipzig [1] and the IAPR International Conference on Machine Learning and Data Mining MLDM [2]. Our intention is not to give a full overview of everything that has been done in the field. We rather want to concentrate on issues that we have been discovering in our work in medical-data analysis and which seems to us to be important for achieving the aim of making intelligent data analysis in medicine a really useful method.

2. The data problem

Medical doctors rely more than other disciplines on common knowledge acquired over the years by a large number of physicians for a special disease. When this common knowledge is lacking, then case studies are reported which can be summarized to common knowledge after a sufficient number of cases has been studied. However, in the case of rare diseases it can take a long time before this knowledge can be induced from the special-case reports.

With the introduction of computers into medical practice and electronic patient records the challenge arises to collect data from different places and to use them for intelligent data analysis. Nevertheless, in order to be able to use this data, a common terminology has to be introduced and the different data bases have to be made compatible. This problem is connected with the topic of ontology and data base design. There has recently been a lot of effort that went into solving this problem. As an example we are reporting here the work done in the EU project INFOGENMED [3,4] and the work done in text mining, in order to learn the terminology/ontology for a domain-specific text collection [5].

Although this is a step in the right direction for the standardization of medical-data collection, there will still be existing a number of different knowledge pieces and data in parallel. Therefore, we still need different types of data-analyzing techniques.

The main problem with medical data and even knowledge is:

- Usually the number of data is limited, depending on the kind of disease. This makes it necessary to collect data from different places and countries in order to get a sufficiently large data collection.
- There will be an imbalanced class distribution. This means that for the one class (usually the class "no disease") there will be a large number of samples, whereas for the other class a small number of samples will be available only (usually for the patient having the disease).
- For some diseases common diagnostic rules have been built up over the years. These rules are available in textbooks or guidelines, but not electronically as expert systems. For creating a good standardization of diagnostic treatment, it is necessary to have expert systems that can guide

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2005.10.003

a medical doctor in his decision. However, rules are not crisp, as there are also some uncertainties included and these uncertainties have to be modeled as well.

3. Methods for intelligent data analysis

The main goal in medical applications is to develop prediction models that can be used for diagnosis. Therefore, most of the papers in this special issue deal with prediction rather than with basic knowledge discovery. A novel method for knowledge modeling is presented by Binchindaritz for phylogenetic classification [6]. New incremental clustering methods that cluster data as long as they arrive in a temporal sequence are developed for hierarchical conceptual clustering by Jänichen and Perner [7] and for partitional clustering by Bouguila and Ziou [8].

A wide variety of methods exist for achieving the goal of prediction. Naïve Bayes classifiers and other statistical classifiers [9] are used as well as methods from machine learning such as decision trees [10,11], case-based reasoning (CBR) [12,13], support-vector machines (SVM) [14,15] or neural nets [16]. Much work has been done to evaluate the different models based on the accuracy on the data of the specific application. This work repeats for any new application. The model used to achieve the desired goal often depends on the application itself, the available data and not seldom from the researcher's preferences. Other facts such as explanation capability, computational time of the model, and data access (incremental or non-incremental) play another role by the selection of the right model.

Researchers studied the influence of feature subset selection on the performance of the prediction models, e.g. such as decision trees [17], and as a result it could be confirmed that feature subset selection is useful for improving the performance of the model. Cohen et al. [18] studied in their paper the influence of the data–sample distribution on different classifiers based on nosocomial infectious-disease data from the University Hospital of Geneva. Their description of the data material shows once more that doing real-world experiments requires a lot of data-preparation work. This work is time-consuming – and not the hot end of research – but it is essential and has to be done carefully in order to obtain good results. The influence of the sample-size distribution is studied on classifiers such as Naïve Bayes, the classifier IB1, decision trees C4.5, Ada Boost and SVM's. They demonstrate in their work once more that a classifier has to be evaluated not only as concerns the overall accuracy, but also with regard to sensitivity and specifity, as well as regarding the class-weighted accuracy [11]. The known random subsampling and oversampling methods are compared to the so-called K-means-based subsampling method and hierarchical-clustering-based subsampling (AHC) proposed by Cohen et al. Their idea is to generate new samples by calculating prototypes from similar cases and using these prototypes to increase/decrease the number of class members. Their study shows that sensitivity and class-weighted accuracy can be significantly improved by a combined strategy of AHC oversampling and K-means subsampling. Using prototypical cases that are created from subclusters within a given class is more reliable than increasing the number of class members by duplicating existing cases. They use knowledge-discovery methods in order to summarize the data in such a way that they suffice for building up a prediction model. Their study is a major contribution towards solving the typical problems in medical diagnosis where positive cases are rare.

Knowledge-based systems are used where knowledge can be obtained from experts during a knowledge-acquisition process. It is clear that this process is time-consuming and needs to be supported by semi-automatic knowledge-acquisition methods [19], especially when uncertain knowledge expressed by membership function and scores has to be modeled. Atzmüller et al. [20] show that these kinds of systems can be supported by machine-learning techniques, in order to obtain the necessary numerical values. Since their method can only handle discrete attributes they use clustering as an unsupervised learning method to discover the intervals for the attribute values. Discretization of numerical attribute values is one important subproblem when building up prediction systems [11]. Supervised and unsupervised methods have been developed [21]. The right discretization of attribute values can significantly improve the performance of a system.

Knowledge-based systems allow one to integrate background knowledge into the reasoning process, so that it is easy to model exceptional rules, which can prevent the system to reason over abnormal conditions. Despite that the result of subspace reasoning can be combined into a fusion rule giving the final diagnosis. In that way they have a similar behavior as decision-fusion systems in pattern recognition [22]. The χ^2 statistics have formerly been applied to learning of the significance of the attributes depending on the class membership and attribute discretization, according to the class

membership in decision-tree induction [11]. Atzmüller et al. [20] make χ^2 statistics applicable to identify dependencies between finding and diagnosing and to learn the diagnostic scores. The proposed strategy has been evaluated carefully within the SONOCONSULT system on 1340 cases.

The summarization of data in a more general form is often not so easy to do, even, when the domain is very complex and has not been intensely studied under knowledge-acquisition terms. However, to make the data immediately available for common use requires techniques such as CBR. CBR can be seen as a method for problem solving as well as for gaining new experience and making it immediately available for problem solving [8]. It can be seen as a learning, knowledge-discovery and knowledge-management method, since it captures from new experience some general knowledge such as case classes, prototypes and some higher level concepts. Montani et al. [23] use CBR to support the treatment of end stage renal-failure patients. They found a way to define a dialysis session, which is mostly a time-series as a case and developed a special-case representation for static and dynamic case features. A pre-selection of a subset of cases is done by classification based on static domain features and afterwards within the class the specific time-series data are considered which are represented by Fourier coefficients obtained with discrete Fourier transformation (DFT). Similarity is determined based on these features, in order to obtain finally the most similar case. Their system is interfaced with the hermodialysis database management unit, in which the patient data are recorded. It can be used by the physician both during a dialysis session and for off-line consultation. With their work they extended the rare work on CBR for one- and two-dimensional signals and it is a pleasure to see that they stepped into this application, which requires overlooking complex knowledge sources comprised of domain-dependent knowledge and signal-processing and signal-representation knowledge.

In basic research in computer vision the application of machine learning and data mining to image segmentation and object recognition has been studied in order to improve the recognition quality. Bhanu et a1. [25] applied genetic algorithm (GA) to the problem of image segmentation, and Perner [26] developed a case-based image segmentation method. These ideas were adopted for the detection of micro-calcification in mammograms by Peng et a1. [27], where with this method it was possible to recognize calcification under varying contrast in the image and varying surrounding tissues. They incorporated a knowledge-discovery mechanism in the genetic algorithm, which could adaptively adjust the fitness values. For the detected objects image features are calculated. Based on a rule set, they evaluate if the recognized spot is a true or false micro-calcification. Decision-tree induction is applied in order to learn these rules and the image features necessary for the evaluation. They follow the same manner as proposed for generic medical image mining in [28] and therefore, their work shows once more that automatic image analysis combined with proper knowledge discovery and machine learning methods yield a flexible and powerful tool for solving various medical image-diagnosis tasks. This has led to a new architecture for image analysis and interpretation systems, for example in high-content analysis of microscopic cell-image interpretation for medical and pharmaceutical applications [29]. Katsis et al. [30] propose a case-based object-recognition and acquisition framework for the classification of individual motor-unit action potentials (MUAPs) from intramuscular electromyographic signals (EMG). Initially, signal pre-processing and candidate MUAP detection takes place. For that the EMG is bandpass-filtered, then the signal is segmented to generate possible waveforms, areas of low activity are eliminated, and peaks as a candidate MUAP waveforms are located and stored. From a test set groups of cases are learnt by using Fuzzy K-Means clustering. From each group of clusters a case representative is learnt and stored into the case base of the matcher. Classification is performed by using support-vector machines. In contrast to [7], no different abstraction levels between the case groups are learnt and the Fuzzy K-Means clustering method is not of incremental fashion. Therefore, the incorporation of new MUAPs requires repeating the whole case-acquisition process again.

Medical doctors usually prefer methods that have an explanation capability so that they can understand why a certain diagnosis has been proposed by the system. The papers presented here in this special issue try to meet this requirement. Depending on the available data methods, such as decision trees, rule-based or case-based reasoning systems are developed. Visualization capabilities have an important role with respect to explanation capabilities. Wagenpfeil et al. [24] developed a tool based on MATLAB for the creation, visualization and interpretation of classical isobolograms in drug–dose response experiments. They show how the results are interpreted based on their application—the potential inhibition of human prostate cancer cell lines with specific retinoids and taxans. The work has been done based on data from the Urologische Klinik of the Technical University in Munich, Klini-

kum rechts der Isaar. Their work shows that classical methods are still useful for medical applications, but that user-friendly and user-supportive software tools are necessary in order to make these methods applicable in clinical practice.

4. Conclusion

The analysis of medical data is a wide and very complex topic. The papers in this special issue use machine learning and intelligent data-analysis methods in order to address special problems in medical data analysis. These problems are: methods to handle imbalanced data, combining machine-learning methods with traditional expert systems, creating CBR systems for multi-modal and complex data sources such as symbolic data and one-dimensional signals and providing useful and practicable software tools with interpretation capabilities for classical data-analysis methods, such as isobolograms. In this respect they advance the state-of-art in medical data analysis. The analysis of their results has been done carefully based on real-world data that had been provided by clinical partners. They provide useful results to the community and give new impulses for further research.

Acknowledgements

We should like to thank all participants of ISMDA 2003 for their contributions to the specific topic, and also the participants of ICDM-Leipzig and MLDM.

References

[1] Industrial Conference on Data Mining, ICDM-Leipzig: http://www.data-mining-forum.de (last accessed 20 October 2005).

[2] IAPR Internal Conference on Machine Learning and Data Mining: http://www.mldm.de (last accessed 20 October 2005).

[3] Maojo V, Garcia-Remesal M, Billhardt H, Crespo J, Martin-Sanchez F, Sousa-Pereira A. A virtual approach to integrating biomedical databases and terminologies. In: Perner P, Brause R, Holzhütter H-G, editors. Medical data analysis, lncs 2868. Berlin: Springer Verlag; 2003. p. 31–8.

[4] INFOGENMED: http://www.infogenmed.net (last accessed: 20 October 2005).

[5] Gillam L, Ahmad K. Pattern mining across domain-specific text collections. In: Perner P, Imiy A, editors. Machine learning and data mining in pattern recognition, lnai 3597. Berlin: Springer Verlag; 2005. p. 570–9.

[6] Bichindaritz I, Potter St. Knowledge based phylogenetic classification mining. In: Perner P, editor. Advances in data mining, applications in image mining, medicine and biotechnology, management and environmental control, and telecommunications, lncs 3275. Berlin: Springer Verlag; 2004. p. 163–72.

[7] Jänichen S, Perner P. Acquisition of concept descriptions by conceptual clustering. In: Perner P, Imiya A, editors. Machine learning and data mining in pattern recognition, lncs 3587. Berlin: Springer Verlag; 2005. p. 153–62.

[8] Bouguila N, Ziou D. MML-based approach for finite Dirichlet mixture estimation and selection. In: Perner P, Imiya A, editors. Machine learning and data mining in pattern recognition, lncs 3587. Berlin: Springer Verlag; 2005. p. 42–51.

[9] Duda RO, Hart PE, Stork DG. Pattern classification New York: Wiley; 2000.

[10] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees Belmont, CA: Wadsworth; 1984.

[11] Perner P. Data mining on multimedia data, lncs 2558 Berlin: Springer Verlag; 2002.

[12] Schmidt R, Gierl L. Temporal abstractions and case-based reasoning for medical course data: two prognostic applications. In: Perner P, editor. Machine learning and data mining in pattern recognition, Lncs 2123. Berlin: Springer Verlag; 2001. p. 23–34.

[13] Marling CR, Petot GJ, Sterling LS. Integrating case-based and rule-based reasoning to meet multiple design constraints. Comput Intell 1999;15(3):308–32.

[14] Li Sh, Fevens T, Krzyzak A, Li S. Automatic clinical image segmentation using pathological modelling, PCA and SVM. In: Perner P, Imiya A, editors. Machine learning and data mining in pattern recognition, lncs 3587. Berlin: Springer Verlag; 2005. p. 314–24.

[15] Karras DA, Mertzios BG, Graveron-Demilly D, van Ormondt D. Improved MRI mining by integrating support vector machine priors in the Bayesian restoration. In: Perner P, Imiya A, editors. Machine learning and data mining in pattern recognition, lncs 3587. Berlin: Springer Verlag; 2005. p. 325–33.

[16] Krzyzak A. Nonlinear function learning and classification using optimal radial basis function networks. In: Perner P, editor. Machine learning and data mining in pattern recognition, lncs 2123. Berlin: Springer Verlag; 2001. p. 217–25.

[17] Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patient's data. In: Perner P, editor. Advances in data mining, lnai 3275. Berlin: Springer Verlag; 2004. p. 153–62.

[18] Cohen G, Hilario M, Sax H, Huyonnet S. Learning from imbalanced data in surveillance of nosocomial infection. Artif Intell Med 2006;37(1):7–18.

[19] Boegl K, Adlassnig KP, Hayashi Y, Rothenfluh TE, Leitich H. Knowledge acquisition in the fuzzy knowledge representation framework of a medical consultation system. Artif Intell Med 2004;30(January (1)):1–26.

[20] Atzmüller M, Baumeister J, Puppe F. Semi-automatic learning of simple diagnostic scores utilizing complexity measures. Artif Intell Med 2006;37(1):19–30.

[21] Perner P, Trautzsch S. Multinterval discretization for decision tree learning. In: Amin A, Dori D, Pudil P, Freeman H, editors. Advances in pattern recognition, LNCS 1451. Berlin: Springer Verlag; 1998. p. 475–82.

[22] Kittler J, Hatef M, Duin RPW, Matas J. "On combining classifiers". IEEE Trans Pattern Anal and Machine Intell 1998;20(3):226–39.

[23] Montani St, Portinale L, Leonardi G, Bellazzi R, Ballazzi R. Case-based retrieval to support the treatment of end stage renal failure patients. Artif Intell Med 2006;37(1):31–42.

[24] Wagenpfeil St, Treiber U, Lehmer A. Statistical Analysis of combined dose effects for experiments with two agents. Artif Intell Med 2006;37(1):65–71.

[25] Bhanu B, Lee S, Ming J. Adaptive image segmentation using a genetic algorithm. IEEE Trans Syst Man Cybern 1995;25(12): 1543–67.
[26] Perner P. An architeture for a CBR image segmentation system. J Eng Appl Artif Intell Eng Appl Artif Intell 1999; 12(6):749–59.
[27] Peng Y, Yao B, Jiang J. Knowledge-discovery incorporated evolutionary search for microcalcification detection in breast cancer diagnosis. Artif Intell Med 2006;37(1):43–53.
[28] Perner P. Image mining: issues, framework, a generic tool and its application to medical-image diagnosis. J Eng Appl Artif Intell;15/2:193–203.
[29] Perner P. Flexible high-content analysis: automatic image analysis and image interpretation of cell pattern. J GIT Imaging Micros 2006:2–3.
[30] Katsis CD, Goletsis Y, Likas A, Fotiadis DI, Sarmas I. A novel method for automated EMG decomposition and MUAP classification. Artif Intell Med 2006;37(1):55–64.

Petra Perner*
Institute of Computer Vision and Applied Computer Sciences, IBal, Körnerstr. 10, 04107 Leipzig, Germany

*Tel.: +49 341 8612 273; fax: +49 341 8612 275
E-mail address: Ibaiperner@aol.com
www.ibai-institute.de